PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-251402
(43)Date of publication of application : 06.09.2002

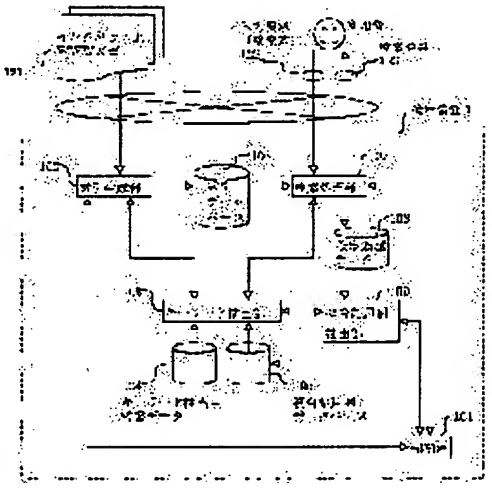
(51)Int.Cl. G06F 17/30
G06F 17/22

(21)Application number : 2001-050257
(22)Date of filing : 26.02.2001

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : AIKAWA TAKEYUKI
SUZUKI KATSUSHI

(54) METHOD AND DEVICE FOR DOCUMENT RETRIEVAL

(57)Abstract:
PROBLEM TO BE SOLVED: To reduce missing of retrieval to be made by a keyword of a word comprising a part of compound words by improving accuracy of analyzing compound words in keyword extraction, which improvement is performed by automatically acquiring a lot of examples of compound words.
SOLUTION: The device extracts keywords from electronic documents at the keyword extracting step and generates the index database at the index generating step, making keywords corresponded to electronic documents. On the other hand, the device generates the database of compound word examples at the compound word extracting step, extracting compound word examples from the record of retrieval made responding to retrieving requests. Further, at the keyword extracting step, the device acquires automatically a lot of compound word examples by extracting keywords using the database of compound word examples and thereby improves accuracy of compound word analysis in extracting keywords and reduces missing of retrieval to be made by a keyword of a word comprising a part of compound words.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－251402
（P2002－251402A）
(43)公開日　平成14年９月６日(2002.9.6)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
Ｇ０６Ｆ	17/30	２１０	Ｇ０６Ｆ 17/30	２１０Ａ	５Ｂ００９
		３３０		３３０Ｃ	５Ｂ０７５
	17/22	５１４	17/22	５１４Ｔ	

審査請求　未請求　請求項の数５　ＯＬ　（全７頁）

(21)出願番号　特願2001－50257(P2001－50257)

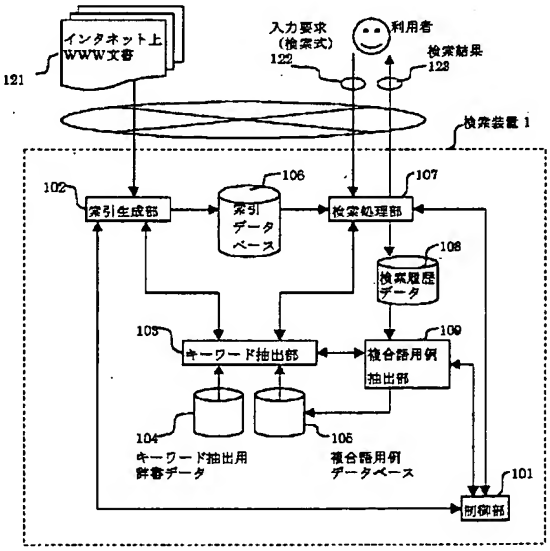
(22)出願日　平成13年２月26日(2001.2.26)

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　相川　勇之
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　鈴木　克志
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　100102439
弁理士　宮田　金雄　（外１名）

Ｆターム(参考)　5B009 MB16 SA12
5B075 NK06 NK14 NK24 NK32 NK39
NR05 NS10 UU06 UU11

(54)【発明の名称】　文書検索方法及び文書検索装置

(57)【要約】
【課題】　従来の検索システムは、形態素解析時に単語区切りの曖昧性があると、誤った索引付けがされる可能性があり、検索漏れを生じる。また、複合語区切りの曖昧性解消手法は、解析誤り修正ルールや用例データを人手で作成するため、大規模な用例データを作成することは困難である。
【解決手段】　キーワード抽出ステップで、電子化文書からキーワードを抽出し、そのキーワードを上記電子化文書に対応付けて索引データベースを索引生成ステップで生成する。一方、複合語用例抽出ステップで検索要求の検索履歴から複合語用例データを抽出して複合語用例データベースを生成し、上記キーワード抽出ステップは、上記複合語用例データベースを用いてキーワード抽出を行なうことで、大量の複合語用例を自動的に収集し、キーワード抽出時の複合語解析精度を向上させ、複合語の一部からなる単語での検索もれを少なくする。



1

【特許請求の範囲】

【請求項1】　電子化文書からキーワードを抽出するキーワード抽出ステップと、抽出されたキーワードを上記電子化文書に対応付けて索引データベースを生成する索引生成ステップと、検索要求からキーワードを抽出して上記索引データベースを検索し、検索結果を作成する検索処理ステップとを備える検索方法において、検索要求を記録した検索履歴から複合語用例データを抽出して複合語用例データベースを生成する複合語用例抽出ステップを有し、上記キーワード抽出ステップでは、上記複合語用例データベースを用いてキーワード抽出を行なうことを特徴とする文書検索方法。

【請求項2】　上記複合語用例抽出ステップは、検索要求を表す検索式が複数の単語を含む場合にこれらを組み合わせた単語入力がなされているかどうかを判定する複合語出現判定ステップを有することを特徴とする請求項1記載の文書検索方法。

【請求項3】　上記複合語用例抽出ステップは、同一ユーザの直前の検索要求を表す検索式に部分文字列となる単語を含む場合に複合語区切りを検出する複合語区切り検出ステップを有することを特徴とする請求項1記載の文書検索方法。

【請求項4】　上記複合語用例抽出ステップは、検索要求を表す検索式の構造を考慮した複合語用例を検索履歴から抽出する手法であることを特徴とする請求項1記載の文書検索方法。

【請求項5】　電子化文書からキーワードを抽出するキーワード抽出部と、このキーワードを上記電子化文書に対応付けた索引データベースを生成する索引生成部と、検索要求からキーワードを抽出して上記索引データベースを検索し、検索結果を作成する検索処理部とを備える文書検索装置において、検索要求を記録した検索履歴から複合語用例を抽出して複合語用例データベースを生成する複合語用例抽出部を有し、上記キーワード抽出部は上記複合語用例データベースを用いてキーワード抽出を行なうことを特徴とする文書検索装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、インターネット上のwww文書や、イントラネット上の電子化文書を検索するための全文検索システムに関するものである。

【0002】

【従来の技術】近年のインターネットの普及および電子化文書の急激な増大に伴い、大規模な検索システムの必要性が大きくなっており、インターネット上のwww文書を検索するための全文検索システムが実用化されている。日本語文書を検索対象とする既存システムのほとんどでは、検索対象文書に対する形態素解析処理を行ない日本語テキストを単語に分割し、これらの単語をキーワードとする索引を生成して検索処理に用いている。

2

【0003】従来システムの一例として、「Goo/InfoBeeが目指す自然言語処理（稲垣他、情報処理学会自然言語処理研究会NL129-4、1999）」（以下、文献1）に開示される技術について図8を参照しながら説明する。従来の検索装置1は、制御部101、索引生成部102、キーワード抽出部103、キーワード抽出用辞書104、索引データベース106、検索処理部107から構成される。

【0004】制御部101では、索引生成部102や検索処理部107の動作を制御する。索引生成部102がインターネット上のwww文書121を収集する時間間隔を制御したり、大量に発せられるユーザからの検索要求を並列処理するための制御を行なう。

【0005】索引生成部102では、検索処理部107における検索処理を高速に行なうための索引データベース106を生成する。インターネット上のwww文書121を収集し、テキスト部分を抽出してキーワード抽出部103においてキーワード抽出用辞書104を参照しつつ形態素解析を行ない、www文書のインターネット上の存在位置を指し示す情報であるURL(Uniform Resource Locator)と、当該文書に含まれるキーワードとを対応づける索引データベース106を生成する。

【0006】検索処理部107ではユーザからの検索要求122を受け付け、検索要求からキーワードを抽出して索引データベース106を検索し、検索結果画面123を生成してユーザ端末に出力する。

【0007】文献1に示されるような形態素解析によるキーワード抽出を用いる検索システムには、単純な文字列検索と異なり単語の区切りを考慮した索引付けがなされるので不要な検索結果が少ないという利点がある。たとえば、文字列検索であれば「義経」を検索しようとすると、「資本主義経済」という検索意図とは異なる検索結果が多数出力されてしまうが、形態素解析によるキーワード抽出では「資本」「主義」「経済」という3つの単語で索引付けされるので、「義経」という検索入力に対して「資本主義経済」が検索されることはない。

【0008】しかし、日本語の単語区切りには曖昧性がある。とくに名詞が連続するような複合語の区切りには、たとえば「現代／用語」と「現／代用／語」のような曖昧性があることが、「日本語形態素解析の誤りの回復について（横尾他、言語処理学会第3回年次大会、pp.429-432）」（文献2）において指摘されている。文献2では、上記のような区切り誤りを手作業で分類して修正ルールを作成し、形態素解析の後処理でこれらの修正ルールを適用することにより区切り誤りを少なくしている。

【0009】また、上記のような曖昧性解消にあたって、区切り情報、品詞情報、意味カテゴリ情報などをあらかじめ人手で付与した用例データを用いた複合語解析手法が、「規則／用例融合型の日本語複合名詞解析法（村中他、言語処理学会第6回年次大会、pp.399-40

2）」（文献３）において提案されている。

【００１０】

【発明が解決しようとする課題】文献１に示される形態素解析に基づくキーワード抽出を行なう検索システムでは、形態素解析の際に単語区切りの曖昧性がある場合は、誤った索引付けがなされる可能性があった。とくに名詞等が連続して出現する複合語の区切り誤りを避けられなかった。そのため、図８のキーワード抽出部103の処理結果によっては、たとえば「現代用語」を含む文書に「現」「代用」「語」という誤った索引付けがされてしまい、「用語」という検索入力に対して検索漏れを生じるという問題があった。

【００１１】文献２および文献３では複合語の区切りの曖昧性を解消する手法が提案されているが、いずれも解析誤り修正ルールや用例データを人手で作成するため作成コストが非常に大きく、インタネット上の大量文書に適用できるような大規模な用例データを作成することが困難であるという課題があった。

【００１２】この発明は上記のような問題点を解決するためになされたもので、ユーザの検索履歴から複合語区切りを自動的に検出して複合語用例データベースに追加する複合語用例データ抽出部を備えることにより、大量の複合語用例を自動的に収集し、キーワード抽出における複合語の解析精度を向上し、複合語の一部からなる単語での検索もれを少なくすることを目的とする。

【００１３】

【課題を解決するための手段】この発明の文書検索方法は、電子化文書からキーワードを抽出するキーワード抽出ステップと、抽出されたキーワードを上記電子化文書に対応付けて索引データベースを生成する索引生成ステップと、検索要求からキーワードを抽出して上記索引データベースを検索し、検索結果を作成する検索処理ステップとを備える検索方法において、検索要求を記録した検索履歴から複合語用例データを抽出して複合語用例データベースを生成する複合語用例抽出ステップを有し、上記キーワード抽出ステップでは、上記複合語用例データベースを用いてキーワード抽出を行なう。

【００１４】また、この発明の文書検索方法は、上記複合語用例抽出ステップが、検索要求を表す検索式が複数の単語を含む場合にこれらを組み合わせた単語入力がなされているかどうかを判定する複合語出現判定ステップを有する。

【００１５】また、この発明の文書検索方法は、上記複合語用例抽出ステップが、同一ユーザの直前の検索要求を表す検索式に部分文字列となる単語を含む場合に複合語区切りを検出する複合語区切り検出ステップを有する。

【００１６】また、この発明の文書検索方法は、上記複合語用例抽出ステップが、検索要求を表す検索式の構造を考慮した複合語用例を検索履歴から抽出する手法からなる。

【００１７】また、この発明の文書検索装置は、電子化文書からキーワードを抽出するキーワード抽出部と、このキーワードを上記電子化文書に対応付けた索引データベースを生成する索引生成部と、検索要求からキーワードを抽出して上記索引データベースを検索し、検索結果を作成する検索処理部とを備える文書検索装置において、検索要求を記録した検索履歴から複合語用例を抽出して複合語用例データベースを生成する複合語用例抽出部を有し、上記キーワード抽出部は上記複合語用例データベースを用いてキーワード抽出を行なう。

【００１８】

【発明の実施の形態】実施の形態1.図１に本発明の実施の形態１．におけるシステム構成図を示す。検索装置１は、制御部101、索引生成部102、キーワード抽出部103、キーワード抽出用辞書104、複合語用例データベース105、索引データベース106、検索処理部107、検索履歴データ108、複合語用例抽出部109から構成される。

【００１９】制御部101では、索引生成部102や検索処理部107、複合語用例抽出部109の動作を制御する。即ち、制御部101は、索引生成部102がインターネット上のWWW文書121を収集する時間間隔を制御したり、大量に発せられるユーザからの検索要求を並列処理するため検索処理部107の制御を行なう。また、複合語用例抽出部109の入力となる検索履歴データ108を出力する。

【００２０】索引生成部102では、検索処理部107における検索処理を高速に行なうための索引データベース106を生成する。インターネット上のWWW文書121を収集し、テキスト部分を抽出してキーワード抽出部103においてキーワード抽出用辞書104および複合用例データベース105を参照しつつ形態素解析を行ない、WWW文書のインターネット上の存在位置を指し示す情報であるURL(Uniform Resource Locator)と、当該文書に含まれるキーワードとを対応づける索引データベース106を生成する。

【００２１】検索処理部107ではユーザからの検索要求122を受け付け、検索要求からキーワードを抽出して索引データベース106を検索し、検索結果画面123を生成してユーザ端末に出力する。このとき、検索処理部107はユーザからの検索要求122の内容を検索履歴データ108に出力する。

【００２２】図２は、図１の複合語用例抽出部109において実行される複合語用例抽出処理の詳細処理フローである。以下、適宜図１およびその他の詳細図面を参照しつつ、図２の各ステップについて説明する。

【００２３】まず、複合語用例抽出部109は図２のステップS201において、検索履歴データ108に含まれるすべての単語を抽出して複合語用例抽出部109の作業用メモリ領域（図示は省略）に格納する。ここで抽出する単語はキーワード抽出部103で形態素解析処理により抽出されるキーワードではなく、実際にユーザが検索要求（検

索式）において記述した単語である。このことについて、図３に示した検索履歴データ108の例を用いて説明する。

【００２４】図３の301は、当該検索要求がどのユーザからいつ発せられたかを示すセッションIDである。この情報はHTTP-cookieなどの既存技術により得ることができるので詳細な説明は割愛する。ここでは以下の詳細処理の説明を簡易にするため、ＩＤの上４桁がユーザ情報を、下４桁が同一ユーザによるセッション情報を表わすものとする。302は、各セッションにおいて入力された検索式である。

【００２５】図３に示した検索履歴データ108において、セッションＩＤが01010001の検索式「現代用語」を図１の検索処理部107で受け付けたときの処理について考える。複合語用例データ105が空の状態では形態素解析処理で区切り誤りの曖昧性を解消できず、「現」「代用」「語」がキーワード抽出部103により抽出される。ステップS201において抽出する単語とは、これらのキーワードではなく、「現代用語」というユーザが入力した単語そのものである。なお、セッションＩＤが00010002の検索式「現代 AND 用語」のように、"AND"や"OR"といった検索用の演算子を含む場合には、これらの演算子を除いた各単語を抽出する。

【００２６】つぎに図３のステップS202に進み、検索履歴108の各検索式について、ステップS203からステップS209の処理を繰り返す。

【００２７】まず、ステップS203では、処理対象の検索式が複数単語を含むかどうかを判定する。たとえば、図３におけるセッションＩＤが00010002の検索式には、「現代」と「用語」の２つの単語が含まれるのでステップS204に進む。複数単語を含まない検索式については、ステップS204からステップS206の処理をスキップしてステップS207に進む。

【００２８】つぎにステップS204では、上記の複数単語を組み合わせて複合語を生成し、この複合語がステップS201において作業用領域に格納した単語に含まれているかどうかを判定する。たとえば、図３におけるセッションＩＤが01010001の検索式には、「現代用語」という単語が含まれるので判定は成功してステップS205に進む。検索式の複数単語を含まない検索式については、ステップS204乃至ステップS206の処理をスキップしてステップS207に進む。

【００２９】なお、上記の例では、「現代 AND 用語」から「現代用語」という複合語を生成して判定を行なったが、複合語の生成にあたって３つ以上の単語から生成する場合には、検索式中の出現順序に従って組み合わせを決定しても良いし、順序を無視してすべての組み合わせを生成しても良い。たとえば「自然 AND 言語 AND処理」という検索式から「自然言語処理」のみを生成しても良いし、「自然言語」と「言語処理」をあわせて生成

してもよい。生成するパターンが増えれば、処理時間がかかるかわりに、獲得できる複合語用例の量が増加する。

【００３０】また、検索式が括弧などにより構造化されている場合には、組み合わせの生成に検索式の構造を反映しても良い。たとえば、「語彙 AND (獲得 OR 抽出)」という検索式の構造を反映して、「語彙獲得」と「語彙抽出」という２つの組み合わせで複合語を生成することも可能である。

【００３１】図２に戻ってステップS205では、ステップS204で抽出された複合語用例が、図１の複合語用例データベース105にすでに登録されているかどうかを判定する。未登録であればステップS206に進み、複合語用例を登録する。既に登録済の用例であれば、ステップS207に進む。上記の例では、「現代／用語」という複合語の用例が登録される。

【００３２】なお、図面を簡易にするため図２には示していないが、ステップS204において複数の用例が抽出された場合には、それぞれの用例についてステップS205の判定を行ない、未登録の用例についてはステップS206で複合語用例を登録する。以下、図４の詳細フローで用例登録処理の内容を説明する。

【００３３】図４のステップS401において、まずステップS204において抽出された用例中の各単語に対して図１のキーワード抽出部103を呼び出してキーワード抽出処理を行なう。これは、生成された複合語の各単語が複数の形態素からなる場合もあるためである。たとえば、「横浜 AND 博物館」という検索式と、「横浜博物館」という検索式から「横浜／博物館」という複合語用例が抽出されたとする。このとき、「博物館」は形態素解析の結果、接尾語の「館」が区切られて「博物／館」となる。このような場合には、「横浜／博物／館」という複合語用例を登録する。

【００３４】つぎに図４のステップS402において、品詞推定処理を行なう。品詞推定では、原則として抽出された複合語用例の右端形態素の品詞を用いる。例えば「現代／用語」の場合は、「用語」という形態素のもつ品詞情報である「名詞」であると推定する。ただし、元の形態素の品詞を変更する働きをもつ例外的な形態素については、これらの形態素が接続した場合の品詞情報をあらかじめ品詞変化一覧表として用意し、これを参照して品詞を推定する。図５に品詞変化一覧表の例を示す。品詞変化一覧表501には上記の働きをもつ例外的な形態素の見出し情報502、品詞情報503、および接続後の複合語がもつ品詞情報504からなる。

【００３５】図６に複合語用例データベースの例を示す。601は見出し情報であり、複合語全体の見出し文字列を格納する。区切り情報602には複合語の区切り位置を格納する。図６ではわかりやすいよう「／」で区切った文字列を用いたが、このような区切り文字を使用する

かわりに分割文字位置を格納することで記憶容量を節減することもできる。品詞情報603は、キーワード抽出処理（後述）において使用する品詞情報である。

【００３６】図２に戻ってステップS207では、同一ユーザの直前の検索式に含まれる単語と、今回の検索式に含まれる単語とで部分文字列関係にあるものを検出する。図３に示した例ではセッションＩＤが02010001の検索式には「参政権」という単語があり、セッションＩＤが02010002の検索式には「外国人参政権」という単語がある。このように前者が後者の部分文字列になっている場合は、ユーザが検索キーワードを長くすることにより検索結果を絞り込もうとした検索履歴であることが推定される。このことから、「外国人／参政権」という区切りが正しいことが推定でき、これを複合語用例として抽出することができる。

【００３７】上記では説明を簡易にするために、連続する２つの検索式が、一方を部分文字列とするそれぞれ単一の検索式となっている例を示したが、複数単語を含む場合も同様に処理可能である。すなわち、「参政権 AND 歴史」という検索式と「外国人参政権 AND 歴史」という検索式が連続して出現すれば、それぞれの検索式に出現する単語の組み合わせのうち、部分文字列となる組み合わせがひとつでも存在すれば、これを抽出すれば良い。また、検索式「端末 OR 通信」と検索式「携帯端末 OR 無線通信」が連続する場合のように、「携帯／端末」と「無線／通信」のように複数の組み合わせを同時に抽出することもできる。

【００３８】また、上記では連続する２つの検索式において、前者が後者の部分文字列になる単語の組み合わせがある場合について説明したが、逆に後者が前者の部分文字列になる単語の組み合わせがある場合についても同様の処理が可能である。たとえば、検索式「違法画像検索」と検索式「画像検索」とが連続して出現したとする。この場合には、ユーザが検索キーワードを短くすることにより検索結果を広げようとした検索履歴であることが推定されるので、上記の「外国人／参政権」の場合と同様に「違法／画像検索」という区切りが正しいことが推定でき、これを複合語用例として抽出することができる。

【００３９】さらに、上記では説明を簡易にするために連続する２つの検索式において部分文字列となる単語の組み合わせでの処理について説明したが、部分文字列となる単語の組み合わせをさがす範囲を広げることも可能である。たとえば同一ユーザの検索式のうち、前後２番目までに含まれる単語で処理することも可能である。また、図３の検索履歴データ108のセッションＩＤに受付処理時刻も含めるようにして、一定時間内に実行された検索式のなかで部分文字列となる単語をさがすということも可能である。

【００４０】図２に戻って、ステップS208では上記のような部分文字列を手がかりにして得られる複合語用例が、図１の複合語用例データベース105にすでに登録されているかどうかを判定する。未登録であればステップS209に進み、複合語用例を登録する。既に登録済の用例であれば、ステップS202に進み次の検索式について処理を続行する。

【００４１】なお、図面を簡易にするため図２には示していないが、ステップS207において複数の用例が抽出された場合には、それぞれの用例についてステップS208の判定を行ない、未登録の用例についてはステップS209で複合語用例を登録する。この点はステップS205およびステップS206と同様である。また、登録の際には、生成された複合語の各単語が複数の形態素からなる場合もあるので、各単語が形態素解析により分割される場合には、分割後の複合語用例を登録するという点についてもステップS206と同様である。以上で図１の複合語用例抽出部109において実行される複合語用例抽出処理の説明を終わる。

【００４２】つぎに図７を参照しながら、図１のキーワード抽出部103において実行されるキーワード抽出処理について説明する。前述のようにキーワード抽出処理には形態素解析処理を用いる。形態素解析のアルゴリズムは良く知られているコスト最小法を用いるものとする。

【００４３】図７のステップS701では、図１のキーワード抽出用辞書データ104を参照して辞書検索処理を行なう。入力された日本語テキストの各文字位置から始まる部分文字列と見出し文字列との照合を行ない、照合に成功した辞書エントリの内容を解析用の作業領域（図示せず）に格納する。キーワード抽出用辞書データ104は、形態素解析用の辞書であり、各形態素の見出し情報や品詞情報を格納している。形態素解析に用いる辞書データおよび辞書検索処理については公知の技術が多数存在するので詳細な処理内容については説明を割愛する。

【００４４】図７のステップS702では、図１の複合語用例データベース105の検索処理を行なう。図６に示した見出し情報601で複合語用例データを検索し、各複合語をステップS701で検索された辞書情報と同様のデータ形式で解析用の作業領域に格納する。ただし、解析結果出力時に図６の区切り情報602を参照できるよう、ステップS701において検索された辞書エントリとは区別できるようなフラグ情報も同時に作業領域に格納する。たとえば、複合語用例データであれば１、ステップS701で検索された辞書エントリであれば０であるとする。

【００４５】図７のステップS703では、コスト最小法アルゴリズムにしたがって解析処理を行なう。コスト最小法では、上記ステップS701およびステップS702において検索された各辞書エントリの品詞情報にしたがって接続検定を行ない、もっともコストの小さくなる接続の組み合わせをコスト最小解として出力するアルゴリズムである。通常、文節数が少なくなるようにコストを設定する

9

と良い解析結果が得られることが経験的に知られている。

【００４６】一般にはステップS701で検索された個別の形態素エントリよりも、ステップS702で検索された複合語用例データのほうが見出しが長いため、上記のように設定されたコストにしたがって解析すれば複合語用例データのほうが優先して解として採用される。ステップS702において作業領域に格納したフラグ情報を用いて、複合語用例データがより優先的に解として採用されるようコストを調整することもできる。

【００４７】図７のステップS704では、ステップS703で求めたコスト最小解を出力する。このとき、作業用領域に格納されたフラグ情報が１であれば複合語用例データなので、図６の区切り情報602を参照して単語分割の結果を出力する。

【００４８】このように、大量の検索要求を記録した検索履歴データから抽出した複合語用例データベースを参照してキーワード抽出を行なうことにより、複合語の区切り誤りが減少する。たとえば「現／代用／語」といった区切り誤りが減少すれば、「用語」という検索要求に対して検索漏れも減少し、好適な検索結果が得られるようになる。

【００４９】以上説明したように、検索履歴から複合語用例を抽出する複合語用例抽出ステップを有することにより、自動的に大量の複合語用例データを抽出できるので、複合語の解析誤りが減少し、複合語を含む文書の検索漏れが少なくなり好適な検索結果が得られるようになる。

10

【００５０】

【発明の効果】以上説明したように、この発明は検索要求を記録した検索履歴のデータから抽出した複合語用例データの複合語用例データベースを生成し、複合語用例データベースを参照してキーワード抽出を行なうことにより、複合語の区切り誤りが減少し、検索漏れも減少して、好適な検索結果が得られるようになる。

【図面の簡単な説明】

【図１】　本発明の実施の形態１におけるシステム構成図。

【図２】　複合語用例抽出処理の詳細処理フロー図。

【図３】　検索履歴データの例を示す説明図。

【図４】　用例登録処理の詳細フロー図。

【図５】　品詞変化一覧表の例を示す説明図。

【図６】　複合語用例データベースの例を示す説明図。

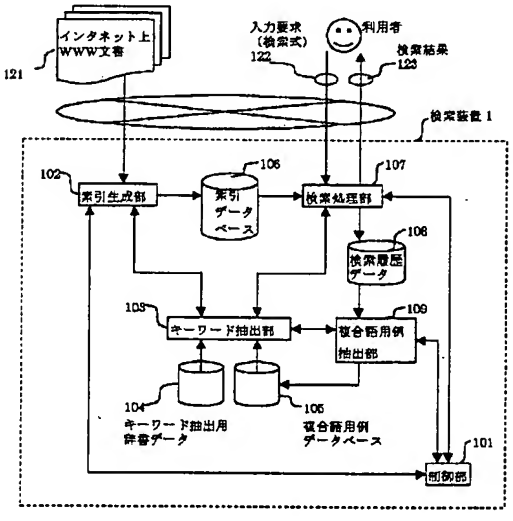
【図７】　キーワード抽出処理の詳細処理フロー図。

【図８】　従来の検索装置のシステム構成図。

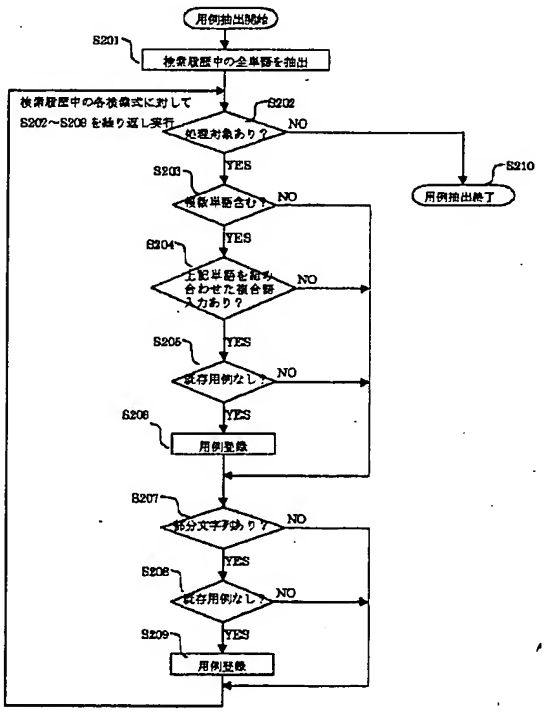
【符号の説明】

１：検索装置、101：制御部、102：索引生成部、103：キーワード抽出部、104：キーワード抽出用辞書、105：複合語用例データベース、106：索引データベース、107：検索処理部、108：検索履歴データ、109複合語用例抽出部。

【図１】



【図２】

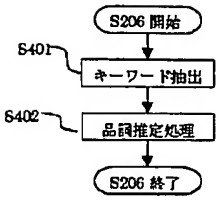


【図３】

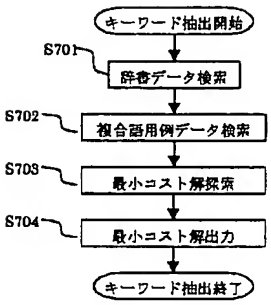
検索履歴データ 108

801 セッションID	802 検索式
00010001	用語
00010002	現代 AND 用語
：	
01010001	現代用語
：	
02010001	参政権
02010002	外国人参政権
：	
03010001	参政権 AND 歴史
03010002	外国人参政権 AND 歴史
：	
04110001	端末 OR 通信
04110002	携帯端末 OR 無線通信
：	

【図４】



【図７】



【図５】

品詞変化一覧表 501

502 見出し	503 品詞情報	504 接続可能形態素の品詞情報
高	接頭語	形容動詞
化	接尾語	サ変名詞
：	：	：

【図６】

複合語用例データベース 105

601 見出し	602 区切り情報	603 品詞情報
現代用語	現代／用語	名詞
横長体物盤	横長／体物／盤	名詞
高価格	高／価格	形容動詞
英文法	英／文法	名詞
高速化	高速／化	サ変名詞
：	：	：

【図８】